Onset Detection in Polyphonic Signals by means of Transient Peak Classification MIREX 2006 – Onset Detection

A. Röbel
IRCAM-CNRS –STMS
1, pl Igor-Stravinsky
75004 Paris, France
axel.roebel(at)ircam.fr

Abstract

The article describes an onset detection algorithm that is based on a classification of spectral peaks into transient and non-transient peaks and a statistical model of the classification results to prevent detection of random transient peaks due to noise. This article describes the algorithmic changes compared to last years submission and discusses the conclusions drawn from the evaluation results.

Keywords: Onset detection. Peak classification.

1. INTRODUCTION

In the following article we are going to describe a transient detection algorithm that has been developed for a special application, the detection of transients to prevent transformation artifacts in phase vocoder based (real time) signal transformations [6, 7]. This application requires a number of special features that distinguishes the proposed algorithm from general case onset detection algorithms: The detection delay should be as short as possible, frequency resolution should be high such that it becomes possible to distinguish spectral peaks that are related to transient and non transient signal components, for proper phase reinitialization the onset detector needs to provide a precise estimate of the location of the steepest ascend of the energy of the attack. In contrast to this constraints the application does not require the detection of soft onsets, where a soft onset is characterized by time constants equal to or above the length of the analysis window. This is due to the fact that such onsets are sufficiently well treated by the standard phase vocoder algorithm. False positive detections are not very problematic as long as they appear in noisy time frequency regions. A major distinction is that a single onset may be (and very often is) composed of multiple transient parts, related either to a slight desynchronization of polyphonic onsets or due to sound made during the preparation of the sound (gliding fingers on a string). While these desynchronized transients are generally not considered as independent onsets they nevertheless constitute transients which should be detected for the intended application.

The evaluation of the transient detection algorithm for onset detection and music segmentation tasks has revealed that the detection results are comparable with existing algorithms for onset detection or signal segmentation tasks [8]. Therefore, it is now the major means for signal segmentation and onset detection in IRCAM's AudioSculpt application [1]. Since MIREX 2005 a number of improvements have been added which should improve the performance with respect to onset detection and which we are interested to evaluate on the MIREX database.

In the following article we will describe the algorithm only briefly, and we refer the reader to the article published during MIREX 2005 [8]. Besides that we will discuss the improvements of the original algorithm since MIREX 2005.

2. Fundamental Strategy

There exist many approaches to detect attack transients. For a number of current approaches see [2, 5, 4, 9]. In contrast to the evaluation of energy evolution in integral frequency bands, a criterion that most of the approaches are relying on, the following article proposes a two stage strategy which first classifies the spectral peaks in a standard DFT spectrum into peaks that potentially may be part of an attack transient and those that are not. Based on this classification a statistical model of background transient peak activity is employed to detect transient events. The advantage of this two stage approach is that the transient components of the signal are classified with rather high frequency resolution, allowing a precise distinction between transient and non transient signal components.

The basic idea of the proposed transient detection scheme is straightforward. A peak is detected as potentially transient whenever the center of gravity (COG) of the time domain energy of the signal related to this peak is at the far right side of the center of the signal window. Note, that it can be shown [8] that the COG of the energy of the time signal and the normalized energy slope are two quantities with qualitatively similar evolution and, therefore, the use of the COG of the energy for transient detection instead of the energy evolution appears to be of minor importance.

3. From transient peaks to onsets

Unfortunately not every spectral peak detected as transient indicates the existence of an onset. Further inspection reveals that spectral peaks related to noise signals quite often have a COG far of the center of the window. In contrast to spectral peaks related to signal onsets these false transient peaks in noise are not synchronized in time with respect to each other. The synchronization of a sufficient number of transient peaks is the final means to avoid detection of noise peaks as onsets.

3.1. Transient energy ratio

Last years MIREX has shown that the normalized energy variation of the transient events is a further means to effectively distinguish random transient events from onsets.

As normalized energy variation (NEV) we define the maximum of the ratio between total signal energy in a transient frame and the transient energy in the same frame, where the maximization is done over the whole duration of the onset. As defined the NEV is bound between 0 and 1. In practical applications the NEV threshold is adapted interactively by the user who can adapt the threshold with direct feedback about the transients that persist. For this years evaluation we used the NEV as control parameter that will be changed for the different runs of the algorithm to create the precision/recall performance curves. We select the NEV threshold to cover the range $NEV = [0 - 0.36]$.

For last years evaluation the NEV was selected to be NEV$= 0.35$. Note, however, that due to the changes of the other parameter settings and the statistical model that was used, the NEV thresholds are not directly comparable.

4. Algorithmic improvements

4.1. Evaluation time grid

A detailed inspection of the algorithm has shown that transient conditions in the statistical model may be limited to only very short time ranges. This is especially true for weak onsets or onsets that appear within a large amount of background noise. An improvement can be easily obtained by means of decreasing the inter frame offset of the analysis frames of the underlying STFT. Compared to MIREX 2005 the frame step has been reduced from an 8th part of the window to an 24th part of the analysis window.

4.2. Limit onset time distance

The results of last years MIREX have shown that one of the major problems of the algorithm are double detections that may occur if multiple instruments have onsets with only slight delays. It is generally desirable to have a means to control the time density of onsets. Accordingly, the algorithm has been changed to allow the user to control the required distance between two detected onsets. If there is more than a single transient event that occurs in the allowed time distance then only the strongest one will be output. The strength of the transients are evaluated according to the NEV criterion described above.

4.3. Transient peak detection

The transient peak detector described in last years MIREX uses the center of gravity (COG) of the energy related to a single peak to determine whether the peak is part of a transient. As explained above, the peak is classified as transient, if the COG is sufficiently to the right of the center. Obviously, for a real transient peak, the signal duration should at the same time be shorter than the duration of the analysis window. Especially for noise peaks it sometimes happens, that the COG is far to the right and the duration is large. This cannot happen in reality because it would indicate that the signal related to the peak would extend outside of the analysis window.

Because the analysis window is fixed, the only way to explain this situation is by means of cancellation. If the part of the signal that lies outside of the analysis window is canceled by other peaks the overall signal stays within the analysis window. This cancellation does happen especially for noise peaks. To detect these transient peak artifacts and to prevent an impact on the transient detector a new mode of the transient peak detector has been developed. This mode requires a transient peak to have an COG offset that is above the user defined COG threshold, and at the same time requires the duration of the signal related to the peak to be smaller then the duration of the analysis window. This mode is enabled in submission 3 of the onset detector algorithm. Note, that the time duration of the signal related to the peak can be calculated directly from the peak spectrum [3].

4.4. Statistical model

The detection of an onset event requires that a sufficient number of synchronized transient peaks are detected. To establish a reasonable condition for the sufficient number we rely on a statistical model of the transient background activity that is due simply to random transient events in the background noise. The background activity is derived by means of a short time history of the detected transient peaks. The history is calculated independently for overlapping bands covering a time range of approximately the 3/4 of the analysis window. For each band the relative number of observed peaks that exceeded the transient threshold is used to determine the average transient probability in the frame history, which in turn is compared to the transient peak probability in the future time range covering approximately 1/4 of the analysis window.

The exact statistical model that is used to describe the transient peak events has been described in [8] and will not repeated here. We address here, the problem of the selection of the bands that are used to monitor the transient events. In the previous version of the algorithm a fixed size band has been used, the bandwidth of which was a priori given by the user. The major problem with the fixed bandwidth of the

Subm. Id	M[ms]	K	G	N_E	H[kHz]	A[dB]	T[ms]	Duration filter
1	36.3	1.7	2.6	15	10	-46	50	off
2	45.4	2.0	2.4	14	11	-46	50	off
3	45.4	1.9	2.4	14	10.5	-46	50	on

Table 1. Parameter settings of the three different submissions to the MIREX 2006 onset detection evaluation. M window size, K COG threshold factor, G transient confidence threshold, N_E minimum bandwidth of statistical model, H upper frequency limit of the spectrum to be used in the detector, A minimum amplitude level of a transient, T minimum distance between two onsets.

statistical model is the fact, that onset events may produce transient peak events with a large scale of different bandwidths. If the bandwidth of the statistical model is much smaller than the bandwidth of the event, the confidence calculated from the model is too small. However, if the bandwidth of the model is much larger than the bandwidth of the event we may not detect a narrowband transient event due to the fact that the variation compared to the background transient activity is too small. To resolve this problem the current version of the algorithm uses a statistical model with different bandwidths. The smallest possible bandwidth is given by the user and the algorithm uses the given bandwidth and all integer multiples of this bandwidths to monitor the background probability. Due to the fact that the models for the larger bandwidths can be calculated from the narrow bands, the calculation of the hierarchic models does not require a significant computational cost. However, it allows us to select the confidence threshold with respect to the optimal bandwidth such that the setup of the threshold is less signal dependent.

5. Parameter selection

There remain a number of user selectable parameters for the transient detector. The first one is the analysis window size M. With respect to this parameter there exist contradicting demands because on one hand attack transients of sinusoids that mix with stationary sinusoids will not be correctly detected such that frequency resolution should be high and window size large. On the other hand we can not detect more than one attack transient within a single window such that window size should be small. This is a variant of the well known time resolution/frequency resolution trade off for time frequency analysis.

The second parameter is the COG threshold. A simple theoretical investigation shows that for the noise free case the maximum COG normalized by the analysis window is 0.5 and for maximum robustness C_s should be close to this value. Due to background noise or preceding notes, however, part of the transient may be covered such that the maximum value of the observed COG will generally be lower than 0.5. As limiting case for a transient condition we consider a linear ramp that start at the very left end of the analysis window. Signals with COG smaller than this will not be detected. The parameter K is a multiplying factor of the COG of the linear ramp and it is used to control the COG threshold. The smaller K the more sensitive the detector is but at the same time the more random transient peaks may be detected in noise.

The third parameter is the confidence threshold G that is the confidence of the statistical model that the transient probability did change. The lower the confidence threshold the more sensitive the algorithm will be, again running the risk of false detections in noise.

The fourth parameter is the minimum bandwidth of the frequency bands that are used to obtain the statistical model for background transient activity. As explained in section **4.4** the statistical model will monitor the transient probability in a hierarchical manner. Therefore, the bandwidth parameter is not as important as in the previous version. The bandwidth N_E is specified in terms of the mainlobe width of the analysis window.

The fifth parameter is the highest frequency H that will possibly be included in the transient peak detection process. The sixth parameter, the minimum distance between detected transients T, has been discussed in section **4.2**. As last parameter we consider the minimum amplitude an onset needs to have to be accepted as onset event. This minimum amplitude A is expressed in terms of the full scale amplitude of the signal.

The parameters have been optimized using a set of hand labeled sound files containing mostly sharp attacks related to drum, bars, or plucked string onsets. Three sets of parameters have been selected to be part of the MIREX evaluation. The parameter settings used in the MIREX evaluation are given in table (4.3).

6. Discussion of the results

While we are happy to see that our algorithm did compare rather favorable with the other contributions, we don't believe that the evaluation can be used to compare the different algorithms.

The main problem is the fact that all algorithms have been trained and adapted using different data sets. Therefore, it appears questionable to draw any conclusions with respect to the ranking of the algorithms. The analysis of the different sound classes reveals that for the different classes different algorithms are “winners”. This could be related to the fact that one algorithm is better, or it could be related to the fact that the algorithm has been trained with a training set that contained more examples for that class. As mentioned above our algorithm has been adapted using especially drums, plucked strings, bells and the like. Accordingly the algorithm should work best with these examples - which is the case. For all these classes the algorithm is rather successful. As a surprise we note that the algorithm works rather successfully well for the singing voice.

Therefore, in the following I will discuss only the ranking of the three versions of our algorithm and the version we send last year.

The first thing to remark is that the recall rate was rather low in all but the bars and bells classes. This indicates that the filtering due to the COG threshold and the confidence level was to strong for some of the data classes. Accordingly, even for a minimum value of the normalized transient energy threshold not all transients passed. This may be due to the fact that our training data base does not contain any sustained strings or wind instruments. We hope to be able to extend the training database to try adapting to a larger scope of sound classes for next years MIREX.

As a second point we note, that the duration filter used in submission three of the algorithm did not always improve the results. The filtering seems to have created an advantage especially for the wind instruments, which may be related to the blowing noise.

As a last remark we mention that the shorter window length of submission 1 seems to be appropriate especially for the voice example. We can only conjecture the reason for this fact. It may be related to vibrato in the singing voice.

Comparing the algorithm with last years version we find that the new versions are better for nearly all classes. Exceptions are the wind and the brass instruments for which the average F-measure have slightly lowered. This result is disturbing especially due to the fact that last years version had a fixed parameter set.

7. Acknowledgments

We would like to thank the team at IMIRSEL for their efforts to create this valuable opportunity to test the algorithms and different parameter settings on an independent data set.

References

[1] N. Bogaards, A. Röbel, and X. Rodet. Sound analysis and processing with audiosculpt 2. In *Proc. Int. Computer Music Conference (ICMC)*, 2004.

[2] J. Bonada. Automatic technique in frequency domain for near-lossless time-scale modification of audio. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 396–399, 2000.

[3] L. Cohen. *Time-frequency analysis*. Signal Processing Series. Prentice Hall, 1995.

[4] C. Duxbury, M. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. In *112th AES Convention*, 2002. Convention Paper 5530.

[5] P. Masri and A. Bateman. Improved modelling of attack transients in music analysis-resynthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 100–103, 1996.

[6] A. Röbel. A new approach to transient processing in the phase vocoder. In *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, pages 344–349, 2003.

[7] A. Röbel. Transient detection and preservation in the phase vocoder. In *Proc. Int. Computer Music Conference (ICMC)*, pages 247–250, 2003.

[8] Axel Röbel. Onset detection in polyphonic signals by means of transient peak classification. In *MIREX Online Proceedings (ISMIR 2005)*, London, Great Britain, September 2005. avail. at `http://www.music-ir.org/evaluation/mirex-results/articles/onset/roebel.pdf`.

[9] X. Rodet and F. Jaillet. Detection and modeling of fast attack transients. In *Proc. Int. Computer Music Conference (ICMC)*, pages 30–33, 2001.